Visualiser des explications contrefactuelles pour des données tabulaires

Victor Guyomard[*,**], François Wallyn[****], Françoise Fessant[*], Thomas Guyet [***]
Tassadit Bouadi[**], Alexandre Termier[**]

[*] Orange, Lannion, France
[**] Univ Rennes, Inria, CNRS, IRISA, Rennes, France
[***] Inria – Lyon, Villeurbanne, France
[****] ENSAI – Rennes, France

Résumé. Dans cet article, nous présentons un outil de visualisation interactif destiné à la visualisation d'explications contrefactuelles. Une explication contrefactuelle se présente sous la forme d'une version modifiée de l'exemple à expliquer qui répond à la question : que faudrait-il changer pour obtenir une prédiction différente ? Ces explications visent à fournir aux utilisateurs des informations personnalisées et exploitables qui leur permettent de comprendre, et éventuellement contester ou améliorer les décisions automatisées. Les résultats sont affichés dans une interface où les explications contrefactuelles sont mises en évidence. Des méthodes interactives sont également fournies pour que les utilisateurs puissent explorer différentes solutions. Le fonctionnement de l'outil est illustré sur un cas d'usage de rétention client. L'outil est compatible avec n'importe quel générateur d'explications contrefactuelles et modèle de décision.

1 Introduction

L'apprentissage automatique est désormais massivement utilisé pour automatiser la prise de décision dans de nombreux domaines, et en particulier dans des domaines qui impactent notre vie quotidienne tels que la santé, le crédit ou encore la justice. Les modèles utilisés sont généralement complexes et opaques. C'est le phénomène de la « boite noire ». L'IA explicable (ou XAI) vise à limiter ce problème en fournissant un ensemble de méthodes pour qu'un utilisateur humain comprenne les facteurs qui ont motivé la décision d'un modèle. L'enjeu de l'explicabilité devient crucial que ce soit pour l'acceptation de l'IA ou le respect des réglementations existantes [1] et à venir [2]. Par exemple, si une personne se voit refuser son crédit, à la suite d'une décision algorithmique, la banque doit être en mesure de lui expliquer les raisons de cette décision. Dans un tel contexte, il pourrait être intéressant de fournir une explication sur ce que cette personne devrait changer pour influencer la décision du modèle.

Les explications contrefactuelles sont un type d'explication permettant d'expliquer la décision du modèle de prédiction à l'aide d'un exemple, proche de l'exemple à expliquer, qui

1. https://gdpr-info.eu/art-22-gdpr/
2. https://artificialintelligenceact.eu/